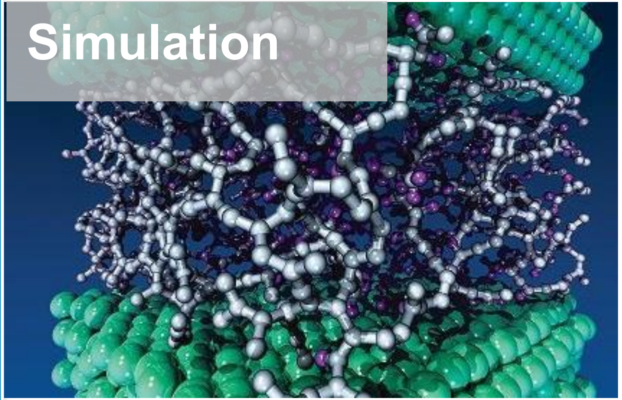
RAPIDS: The SciDAC Institute for Computer Science and Data

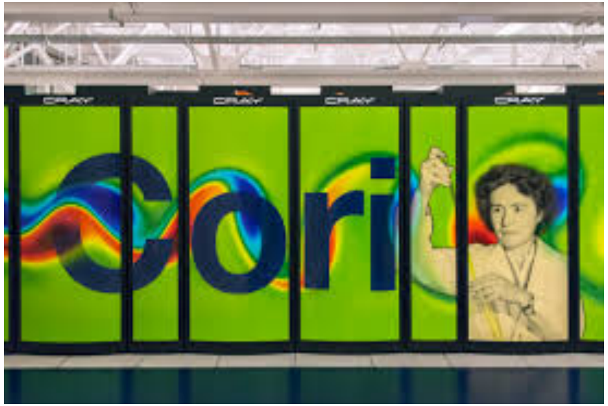
ROBERT ROSS
Institute Director
Argonne National Laboratory
rross@mcs.anl.gov

LENNY OLIKER
Deputy Director
Lawrence Berkeley National Laboratory
loliker@lbl.gov

Diverse Science and Systems

RAPIDS



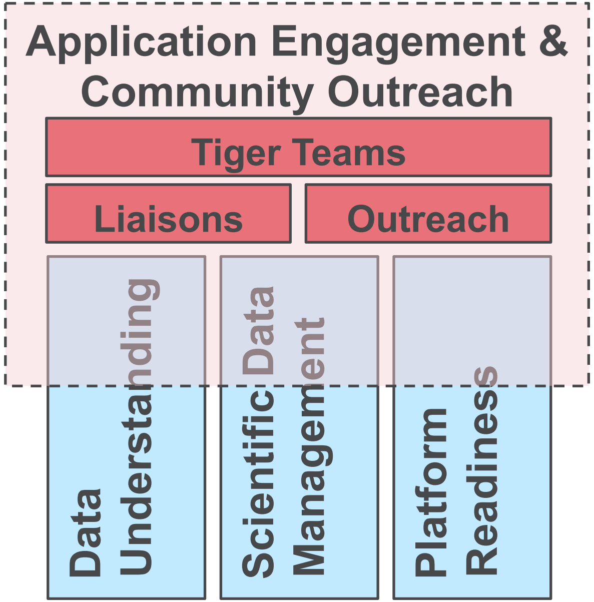


Top image credit B. Helland (ASCR). Bottom left, center, and right images credit ALCF, NERSC, and OLCF respectively.

The RAPIDS Institute

Solving computer science and data challenges for Office of Science application teams to achieve science breakthroughs on DOE platforms.

- **Technology Focus Areas**
  - **Data Understanding** – scalable methods, robust infrastructure, machine learning
  - **Scientific Data Management** – I/O libraries, coupling, knowledge management
  - **Platform Readiness** – hybrid programming, deep memory hierarchy, autotuning, correctness
- **Application Engagement**
  - *Tiger Teams* engage experts in multiple areas
  - Software productivity: verification and validation, etc.
  - Outreach activities connect with broader community



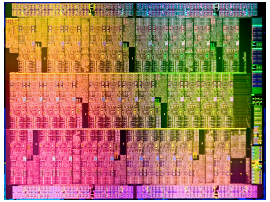
Argonne National Laboratory · Berkeley Lab · University of Delaware · Lawrence Livermore National Laboratory · Oak Ridge National Laboratory · University of Oregon · Rutgers · Brookhaven National Laboratory · The HDF Group · Kitware · Los Alamos National Laboratory · Northwestern University · The Ohio State University · The University of Utah

Platform Readiness

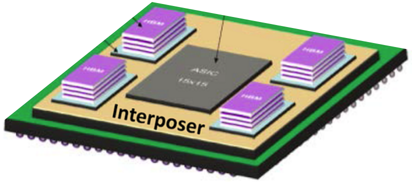
JEFFREY VETTER
ORNL

PAUL HOVLAND
ANL

Exascale: End of the CMOS Era



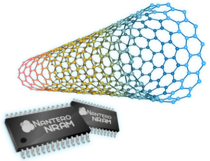
O(100 TF) per node
Wide vectors / GPUs
More for specialized procs



Memory system
on package

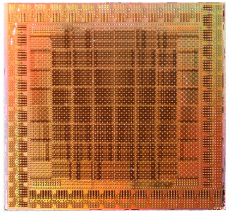
Disk for
archive

Flash for
main storage



NVM for
storage cache

O(100k) nodes, ~30-50 MW
O(10 Exaflops)



Low diameter networks with optics

Over the next decade, computers won't change that much from the current model.

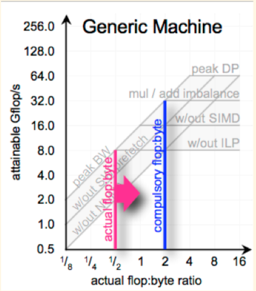
S. Scott, "Beyond Exascale: Playing the CMOS Endgame," SOS23, March 2019.

Platform Readiness

Preparing scientific codes for current and upcoming system through application of best-in-class expertise and tools.

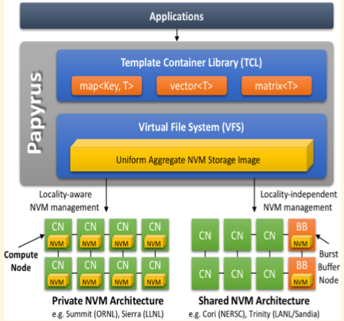
Performance Modeling/Analysis

- TAU: Performance Analytics & Tuning for Heterogeneous HPC
- Roofline: Easy-to-understand, visual performance model



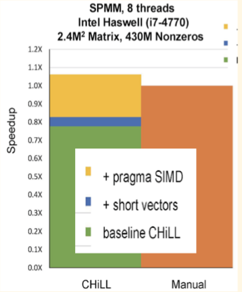
Portable Programming

- For heterogeneous systems, deep memory hierarchies
- Papyrus: abstractions for shared data using map, vector, and matrix modalities



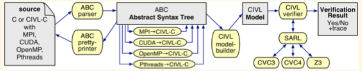
Code Generation and Autotuning

- CHILL: model-based Autotuning
- Sweeps optimization parameters for target platform
- Enables effective use of accelerators without multiple code versions



Program Correctness

- CIVIL: Static verification of HPC programs
- Uses static analysis techniques over well-defined input ranges to do symbolic execution
- Enables Verification equivalence of two implementations

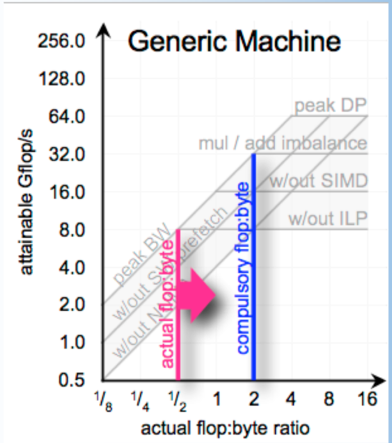


Roofline Performance Modeling

ASCR Base & LDRD

Developed Roofline concept
2006-2011:

- Easy-to-understand, visual performance model
- Offers insights to programmers and architects on improving parallel software and hardware.

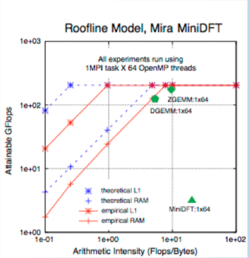


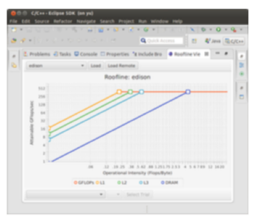
Proof of concept successfully applied to numerous computational kernels and emerging computing systems.

SciDAC3 Development

Roofline augmentation under SciDAC3
2013-2017:

- Collaboration with FASTMath SciDAC Institute
- Developed Empirical Roofline Toolkit (ERT) with public release 03/2015, with Roofline Visualizer
- Created community tool for automatic hardware introspection and analysis

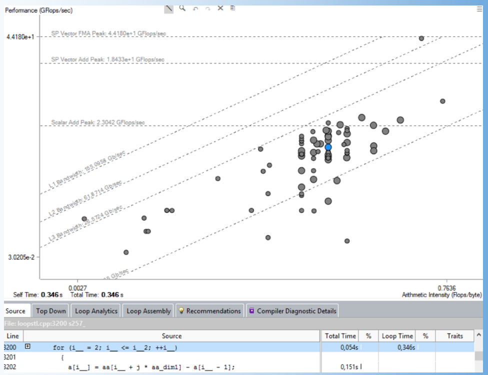




Automated Roofline code used to diagnose performance problems for DOE and SciDAC codes.

Outcome & Impact

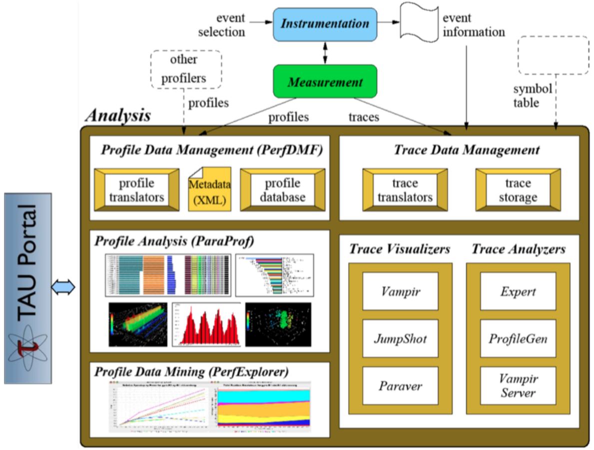
- Roofline has become a broadly used performance modeling methodology across DOE
- Intel has embraced the approach and integrated it into its production Intel® Advisor
- **Collaboration with NERSC to instrument and analyze execution of real applications on machines such as Edison and Cori**

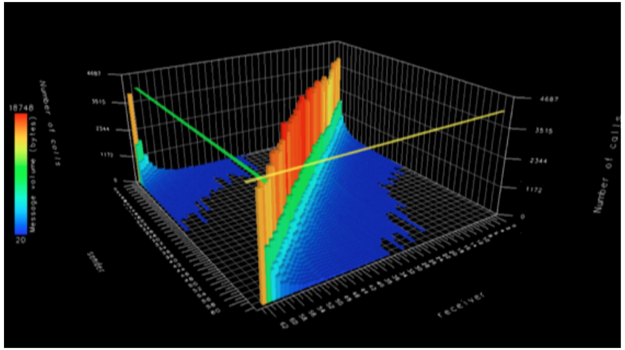


Snapshot of existing Intel Roofline tool in practice.

Performance Observation, Analytics, and Tuning for Heterogeneous Platforms with TAU

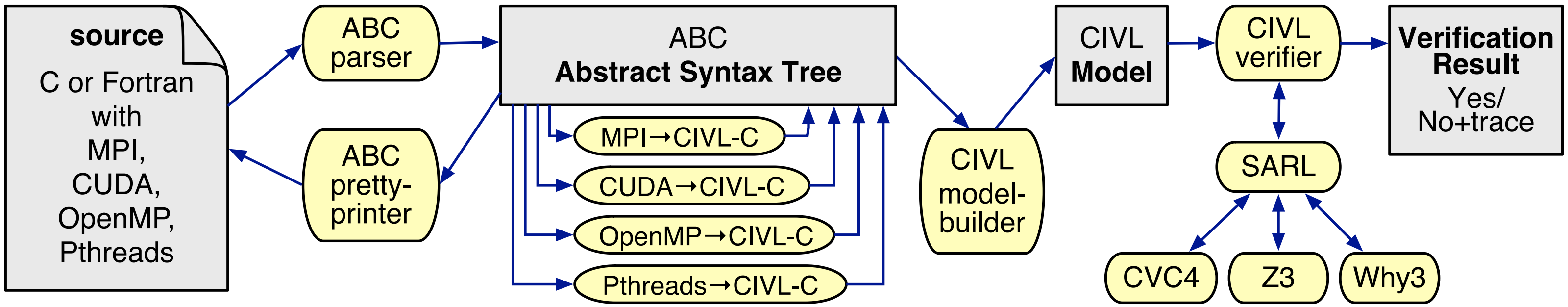
- Heterogeneous software stacks
  - Languages: OpenMP, OpenACC, CUDA, ROCm
  - Libraries/Metaprogramming: Kokkos, RAJA
  - Hybrid: MPI+X
- Runtimes
  - OpenMP, MPI, I/O, asynchronous multitasking
- Compilers and autotuners
  - LLVM, Chill, Oreo, Active Harmony, OpenARC
- Heterogeneous hardware measurement
  - Memory, Power, Network
- Integration with ADIOS2 for both I/O library measurement and ADIOS2 output of application performance data





Static Verification of HPC Programs (CIVL)

RAPIDS



- Source may include CIVL-C primitives: **input**, **output**, **assert**, **assume, …**
- All concurrency translated to **CIVL-C**
- **Fortran** and **C** translated to same intermediate language, CIVL-C
- Program may be composed of multiple translation units (including Fortran+C)

- Verifier uses **symbolic execution** to check properties for all possible inputs (within specified bounds)
- Absence of: assertion violations, deadlock, illegal pointer operations, out-of-bound indexes, OpenMP data-races, …
- Verify **equivalence** of 2 implementations

Reproducible Performance Analysis
with HEP and NUCLEI

RAPIDS

Scientific Motivation

Develop a data analytics platform for creating and reusing the performance analysis workflows for improving the performance of DOE science codes.

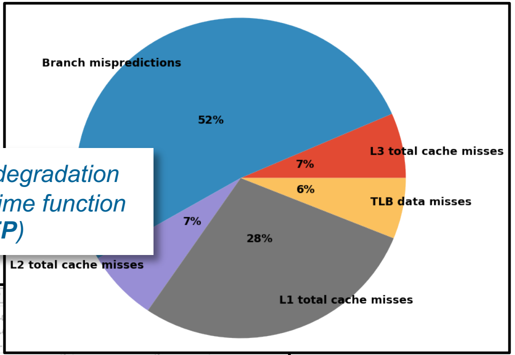
- **HEP Event Tracking:** Effective utilization of many-core SIMD and SIMT
- **NUCLEI:** Exploiting hybrid distributed- and shared-memory parallelism in integrated legacy and newly developed codes

Significance and Impact

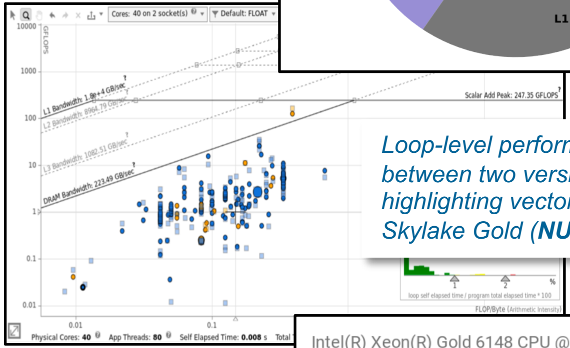
By enabling customizable, reusable performance analysis that can be maintained and extended by application teams, we can reduce the reliance on expert help and speed up performance optimization.

Research Details

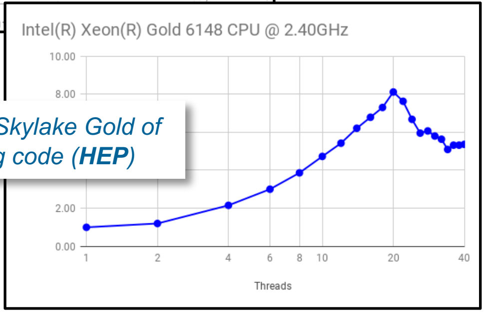
- Parallelized Kalman filter tracking, mkFit (**HEP Event Tracking**): Used TAU Commander and Python Pandas to create performance analysis “recipes”. Achieved 2.7x speedup from explicit vectorization and > 10x from shared-memory parallelization on KNL; 4.4x speedup when integrated into main CMSSW framework (without optimizing data conversion).
- HFODD (**NUCLEI**): Used TAU and Intel’s VTune and Advisor tools to create automated analysis workflows for shared-memory scaling and vectorization.



*Sources of performance degradation in the highest execution-time function in mkFit on Intel KNL (**HEP**)*



*Loop-level performance comparison between two versions of HFODD, highlighting vectorization on Intel Skylake Gold (**NUCLEI**)*



*Strong scaling on Skylake Gold of OpenMP hit finding code (**HEP**)*

Lawrence Livermore National Laboratory
Fermilab
UNIVERSITY OF OREGON

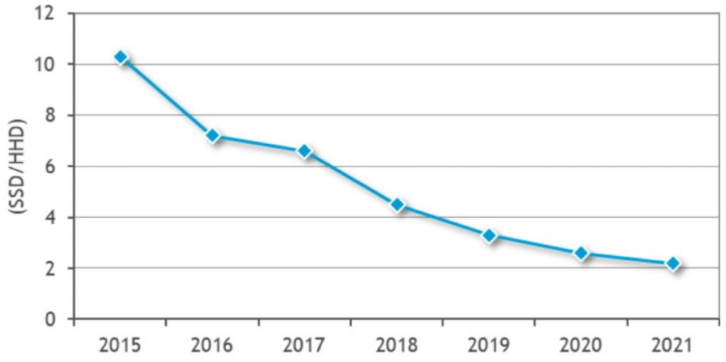
Data Management

SCOTT KLASKY
ORNL

JOHN WU
LBNL

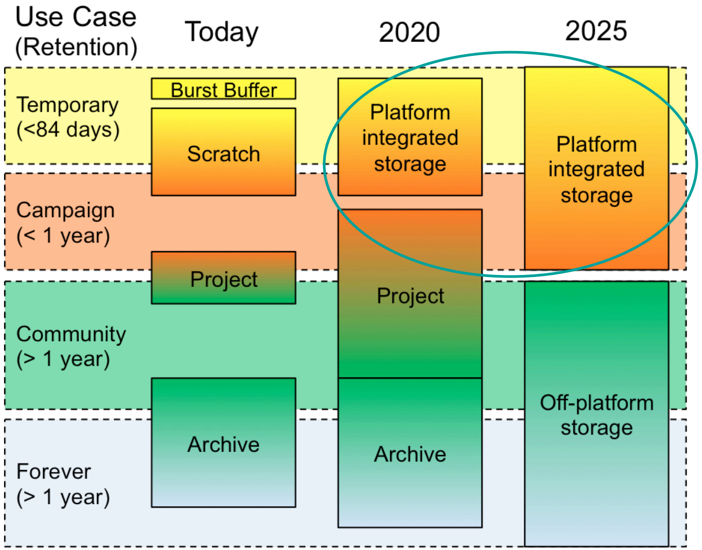
Technology Change in Storage Systems

Solid-state disk vs. hard disk drive pricing (per GB ratio)



Source: Hyperion research
https://www.storagenewsletter.com/2018/08/07/flash-storage-trends-and-impacts

Evolution of the NERSC storage hierarchy between today and 2025



Continued decline in cost of SSD capacity relative to HDD has led to plans to employ SSD-backed platform storage, integrated into the platform.

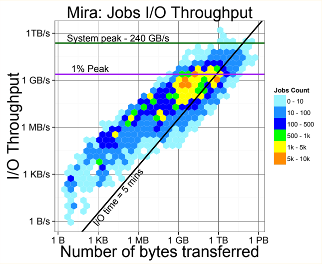
G. Lockwood et al. "Storage 2020: A Vision for the Future of HPC Storage," October 2017, https://escholarship.org/uc/item/744479dp

Data Management

Deploying and supporting efficient methods to move and manage data in a scientific campaign.

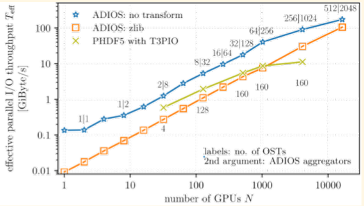
Performance Monitoring

- Understanding of I/O performance at scale
- Darshan: “Always on” statistics gathering
- TAU: Fine-grained I/O tracing of operations at multiple layers



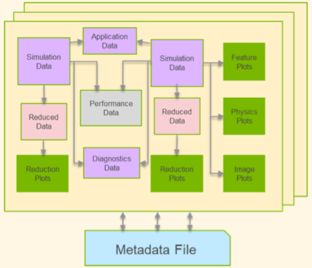
Storage and I/O

- HDF5: A data model, parallel I/O library, and file format for storing and managing data
- Parallel netCDF: Provides parallel access to traditional netCDF datasets
- ADIOS: community I/O framework to enable scientific discovery



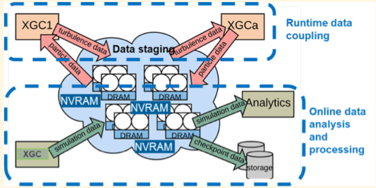
Knowledge Management

- FastBit: Organize and quickly find records across files generated and used during a scientific campaign
- Manage and query the data across a scientific campaign



Code Coupling

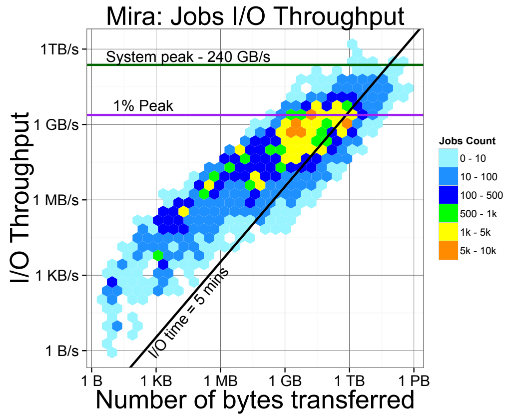
- Dataspaces & SST2: In-memory storage distributed across set of cores/nodes, using RAM and/or NVRAM
- Fast I/O to couple codes together asynchronously
- In-staging data processing, querying, sharing, and exchange

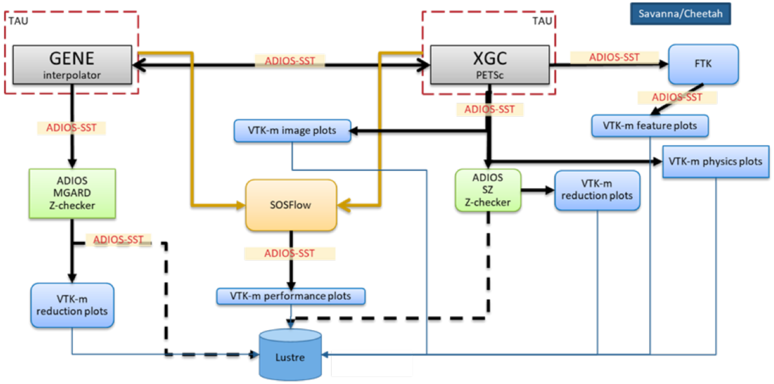


Performance Monitoring

Enabling understanding of I/O performance at scale

- Darshan
  - **"Always on" statistics gathering**
  - Observes I/O patterns of applications running on production HPC platforms, without perturbing execution, with enough detail to gain insight and aid in performance debugging
- TAU
  - **Fine-grained tracing of I/O operations at multiple layers**
  - ADIOS2 integration: integrated profile instrumentation of ADIOS2 and ability to stream TAU application performance data directly out to ADIOS2 at runtime

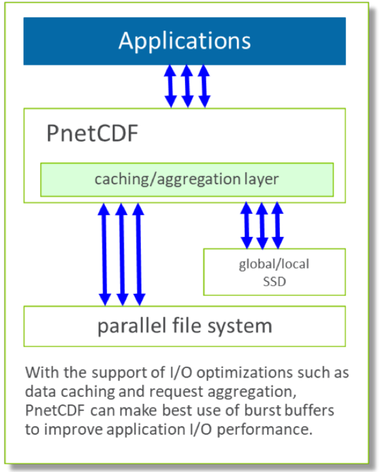




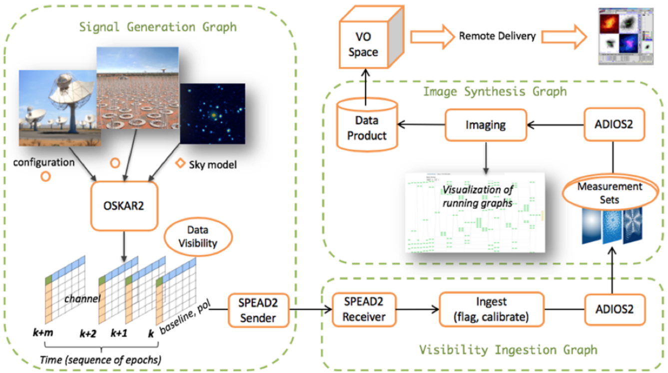
Storage and I/O

Libraries/frameworks to assist in fast and portable I/O

- **HDF5**
  - A data model, parallel I/O library, and file format for storing and managing data
  - Flexible, self-describing, portable, high performance
- **Parallel netCDF**
  - Provides parallel access to traditional netCDF datasets
  - Includes algorithms for accelerating common patterns such as multi-variable writes
- **ADIOS**
  - A community I/O framework to enable scientific discovery
  - In-memory code coupling for applications to other applications and/or analysis/visualization
  - Incorporates the state of the art I/O techniques for checkpoint, self describing data, and in situ data movement between codes

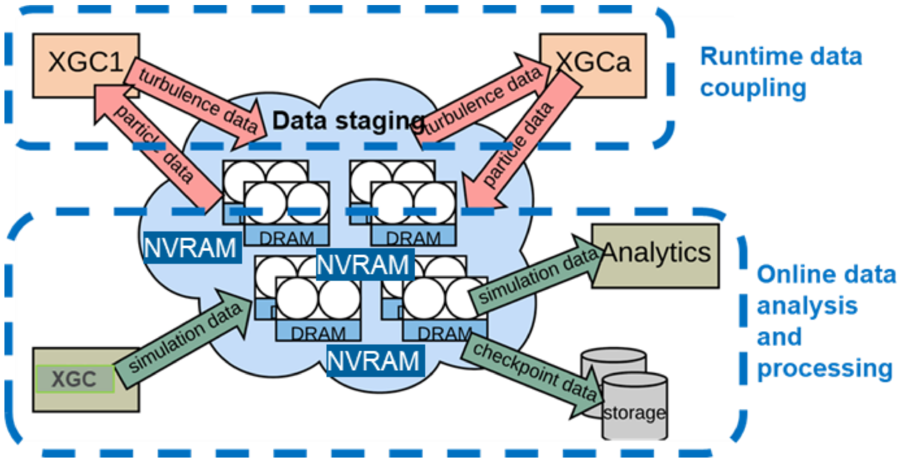


ADIOS is used for the backend for SKA data movement/storage



Code Coupling: DataSpaces

In-memory storage distributed across set of cores/nodes, using RAM and/or NVRAM



The figure shows an in-situ fusion simulation workflow with code coupling and in-situ data processing. DataSpaces provides a semantically specialized shared-space abstraction using staging resources to support dynamic and asynchronous coordination, interactions, and data exchanges between components of an in-situ workflow.

- Fast I/O to asynchronously couple codes together
- Couple simulation, visualization, analysis, and performance monitoring
- In-staging data processing, querying, sharing, and exchange
  - Virtual shared-space programming abstraction
  - Provides an efficient, high-throughput/low-latency asynchronous data transport
  - Predictive data movement and layout

Global Particle-in-Cell Simulation of Fusion Plasmas

Scientific Achievement

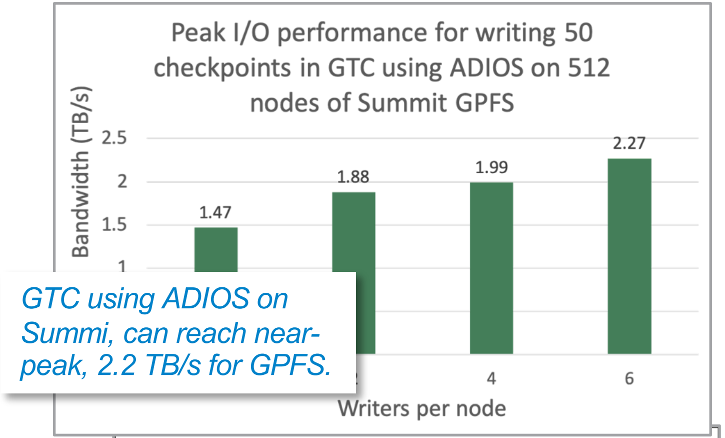
Energetic particle (EP) confinement is a key physics issue for the burning plasma experiment ITER. By enabling GTC with the ADIOS framework, we can finally write the majority of the physics data with minimal impact on the code performance on the Summit HPC resource at the OLCF

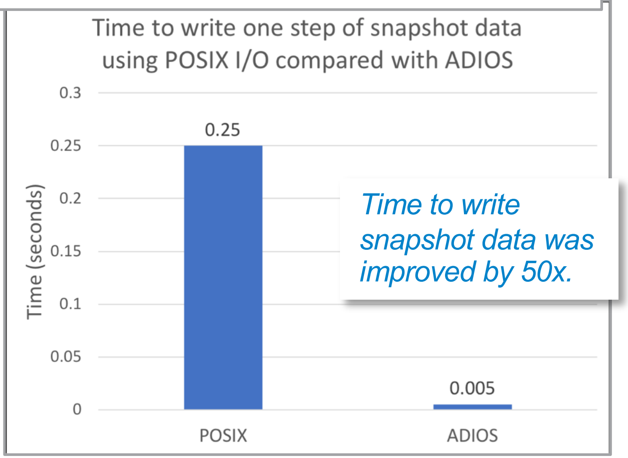
Significance and Impact

- GTC can generate data, over 100 TB of physics data every hour
- GTC has been equipped with ADIOS to allow all of the relevant physics information to be written to the Summit GPFS file system in less than 3% of the total runtime
- New data analytics is being written for GTC to work in both post-processing and in situ workflows

Research Details

- A new “engine” inside of ADIOS was developed to allow for extreme performance for Particle In Cell code I/O





L. Wan, K. Mehta, et al, “Data Management Challenges of Exascale Scientific Simulations: A Case Study with GTC and ADIOS”, ICCM 2019 (accepted)

U.S. DEPARTMENT OF ENERGY | Office of Science

RAPIDS

Data Understanding

DMITRIY MOROZOV
LBNL

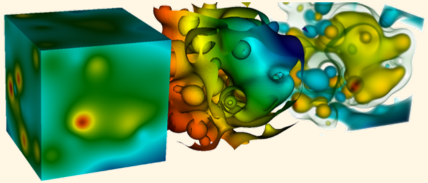
PRASANNA BALAPRAKASH
ANL

Data Understanding

Facilitating understanding of large and complex science data through robust and scalable analysis methods, including learning approaches.

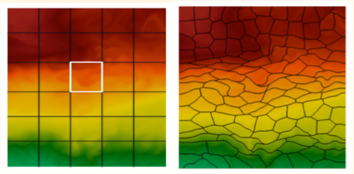
Visualization

- Visualization tools that leverage modern HPC
- In situ frameworks, to enable efficient system usage
- Scalable infrastructure: service oriented data analysis and reduction
- Leveraging deep memory hierarchy, on-node parallelism
- Analysis/visualization of high dimensional datasets



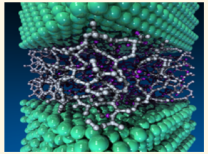
Scientific Data Analysis

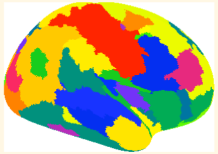
- Feature detection for visualizing and comparative analysis
- Geometric analysis: Delaunay/Voronoi tessellation
- Statistical analysis of ensemble and uncertain data
- Uncertain flows from ensemble modeling
- Topological features in scalar fields



Machine Learning

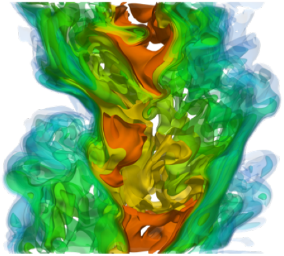
- Supervised learning methods, including deep learning for object classification
- Unsupervised learning methods, including dimension reduction
- Scalable parallel graph algorithms
- Sparse inverse covariance matrix estimation



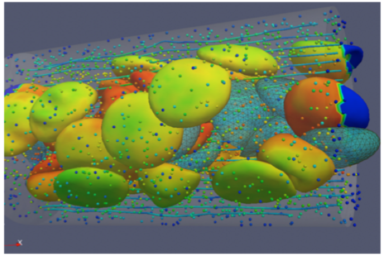


Visualization

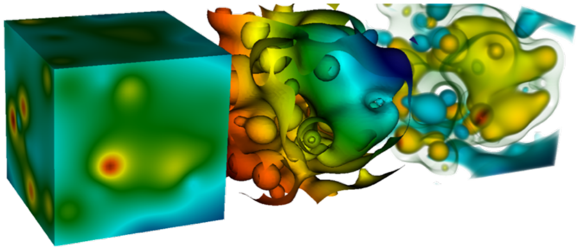
Feature-rich visualization tools that can be run at scale, in situ



- Successful existing tools: **ParaView** and **VisIt,** both built on top of **VTK**, take advantage of massively parallel architectures of modern super-computers
- In situ frameworks, **VisIt/libsim, ParaView/Catalyst, ADIOS, Sensei**, enable using these systems efficiently with the simulations, e.g., to visualize live simulations avoiding the IO bottleneck



- **Scalable infrastructure**: service-oriented data analysis and reduction, co-analysis with performance data
- Major focus on adapting to the deep memory hierarchies and massive on-node hybrid parallelism **(VTK-m)**
- Also useful information visualization techniques **(EDEN)**, techniques for analysis and visualization of high-dimensional datasets

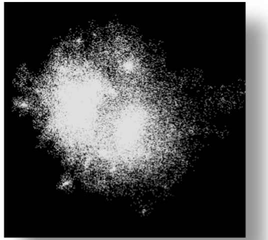


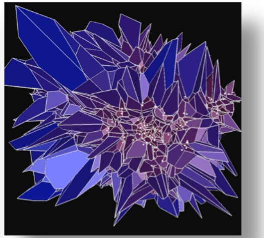
VisIt VTK ParaView VTK-m

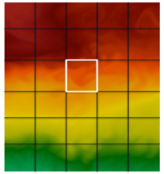
Scientific Data Analysis

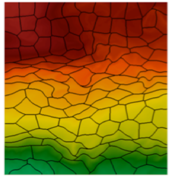
Scalable methods for finding and analyzing features of importance

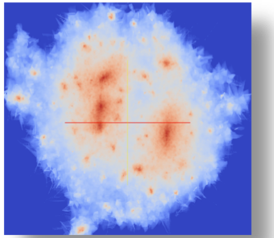
- Expertise in **feature detection**, traditionally for visualization and comparative analysis. Moving forward as input to machine learning methods.
- **Geometric analysis (tess):** scalable computation of Delaunay and Voronoi tessellations, e.g., for density estimation in cosmological data
- Statistical analysis of ensemble data **(edda)**:
  - representation of large scale uncertain data
  - analysis of ensemble and uncertain features
  - exploration of parameter space for ensemble simulations
- **Uncertain flows** from ensemble modeling (fluid dynamics, climate, weather)
  - Generalizing flow features for uncertain data
  - Surface Density Estimates to quantify uncertainty
  - Scalable algorithms to stochastically trace particles
- **Topological features** in scalar fields
  - Scalable computation of merge trees, contour trees, persistence diagrams (used in cosmology, combustion, materials science, etc.)
  - Useful both for visualization and for comparison of simulations, to each other and to experiments

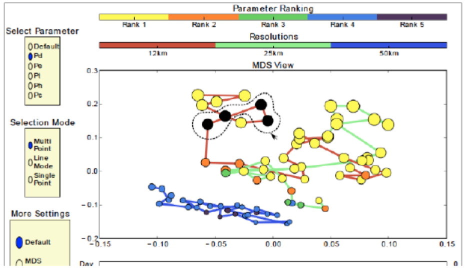












Statistical Super Resolutions for Large Scale Ensemble Cosmological Simulations

RAPIDS

Scientific Achievement

Enable scientists to reduce the storage space requirement when running large ensemble simulations, while still make it possible to perform full scale simulation parameter exploration for post-hoc analysis

Significance and Impact

With the statistical signatures, it is now possible to reconstruct simulation output of novel parameters that was not saved during simulations. The space saving can be more than 95% as compared to saving compressed results of all runs.

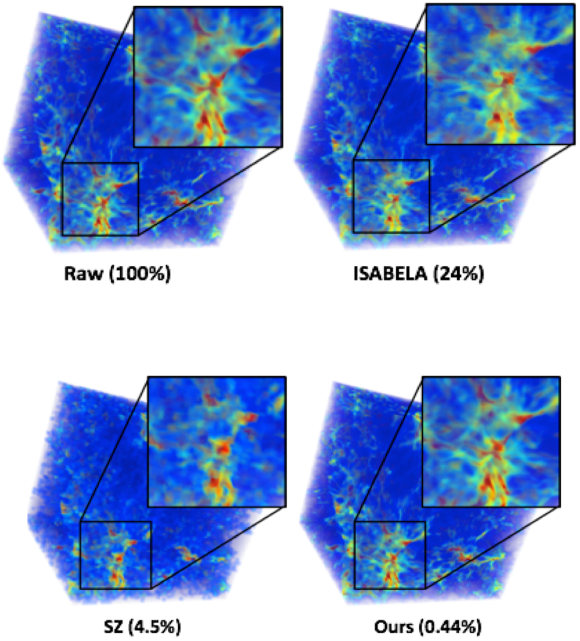
Research Details

– Store a small number of simulation results at full resolution into a code book as prior knowledge

– Down sample remaining data into GMMs as the statistical signatures

– Data at an arbitrary parameter configuration can be reconstructed from the prior knowledge and the statistical signatures

– The priori knowledge only takes 0.44% of the original data for a cosmology simulation using Nyx



Images produced by our super resolution representations

Los Alamos NATIONAL LABORATORY
Argonne NATIONAL LABORATORY
Extreme-scale Distribution-based Data Analysis
THE OHIO STATE UNIVERSITY

Supporting New Science Communication Patterns and Data Models

RAPIDS

Scientific Achievement

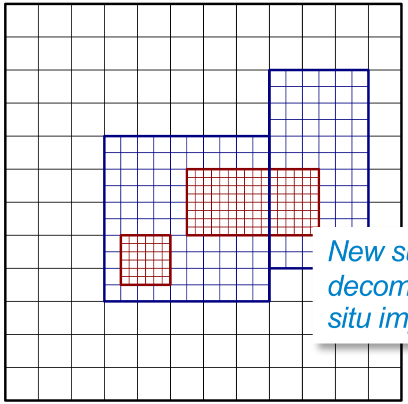
New communication algorithms in DIY enable efficient implementation of more advanced analysis algorithms and support for working in situ with advanced simulation data models (AMR).

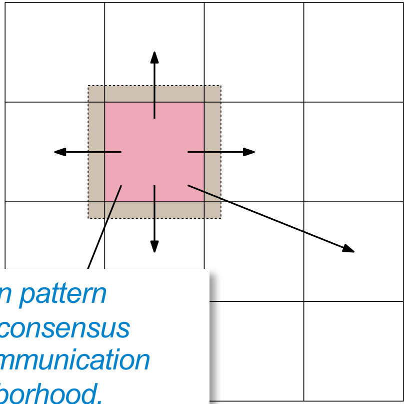
Significance and Impact

DIY provides distributed programming support for VTK-m as well as standalone analysis algorithms (including HEPonHPC and CANGA partnerships).

Research Details

– Added distributed consensus protocol (rexchange) that enables efficient communication between arbitrary pairs of blocks.
– Added support for AMR data, enabling in situ analysis of advanced simulations.
– Better integration with VTK-m and ParaView (Kitware).





Argonne NATIONAL LABORATORY
Lawrence Berkeley National Laboratory

Machine Learning and AI

Executive Order 13859 of February 11, 2019

Maintaining American Leadership in Artificial Intelligence

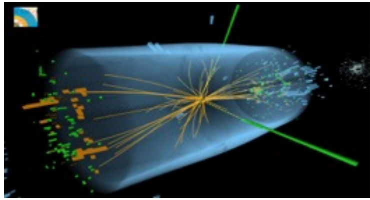
By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

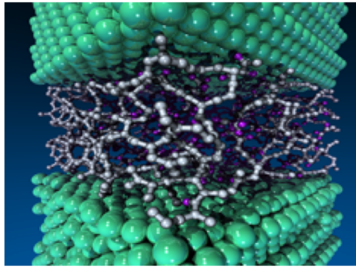
Section 1. *Policy and Principles.* Artificial Intelligence (AI) promises to drive growth of the United States economy, enhance our economic and national security, and improve our quality of life. The United States is the world leader in AI research and development (R&D) and deployment. Continued American leadership in AI is of paramount importance to maintaining the economic and national security of the United States and to shaping the global evolution of AI in a manner consistent with our Nation's values, policies, and priorities.

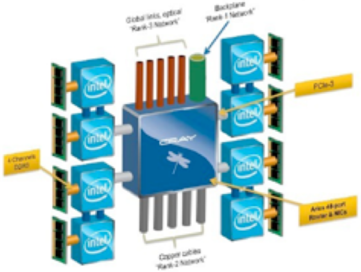
Machine Learning

Domain-specific applications of deep learning, predictive performance models, data- and model-parallel training

- Supervised learning methods:
  - Deep learning for object classification and identification
  - Automatic multiobjective modeling **(AutoMOMML)** to simplify model selection
  - Asynchronous hyper-parameter and neural arch. search **(DeepHyper tools)**
  - Autotuning parameters for code/application **(SuRF)**
  - Performance, power, and energy modeling of novel HPC architectures;
- Unsupervised learning methods:
  - Manifold learning/dimensionality reduction; approximation algorithms for streaming data, streaming spectral clustering
  - Useful for adaptive sampling (e.g., for molecular dynamics trajectories)
- Reinforcement learning
- Scalable parallel **graph algorithms (LAGraph)**:
  - Recast graph algorithms into linear algebra operations
  - Building blocks and communication-avoiding algorithms applied to neural nets
- Tools for understanding ML models **(DeepVid, GANViz, DQNViz)**







Using Roofline to Characterize TensorFlow on GPUs

Scientific Achievement

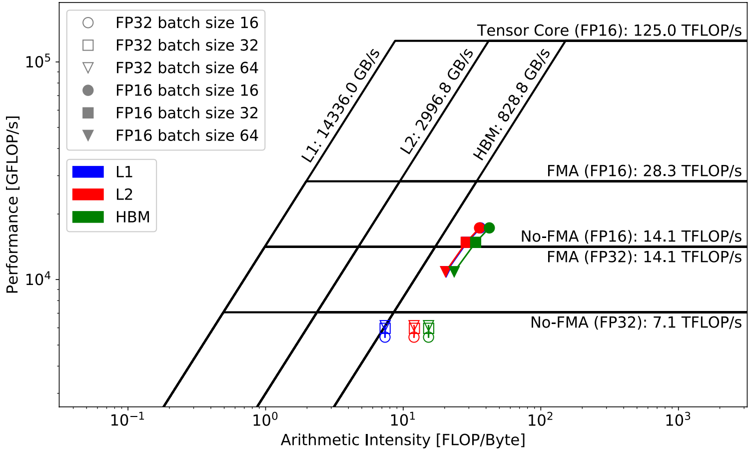
Created a methodology for analyzing the execution of GPU Tensor Core-accelerated DL/AI applications using Roofline.

Significance and Impact

This work enables Roofline-based analysis of NVIDIA Tensor Core accelerated AI/DL applications including quantitative assessments of TensorFlow performance on NVIDIA Volta GPUs.

Research Details

- Collaboration between RAPIDS, NERSC, and NVIDIA
- Formulated methodology for using NVProf to analyze tensor-core accelerated applications using Roofline
- Used Roofline to analyze the forward and backward phases in TensorFlow as a function of FP16 and FP32.
- TensorFlow cannot sustain the theoretical 125TF/s due to a lack of locality and data permutation overheads.



TensorFlow (forward pass) on V100

Results shown are relative to precision (32b and 16b tensor cores) and batch size (16,32,64). Although tensor cores deliver >2x performance, performance is far from theoretical 125TF/s

Yang et al., “Hierarchical Roofline Analysis for GPUs: Accelerating Performance Optimization for the NERSC-9 Perlmutter System”, CUG, 2019.

Understanding How Deep Learning Models Operate

Scientific Achievement

Allow developers of deep learning models to open the black box to see how and why the DNN model functions, so as to further optimize its performance

Significance and Impact

Explaining AI decision-making is a key challenge in the adoption of AI algorithms in scientific activities. Visual analytics approaches can play a crucial role in explaining modern AI models.

Research Details

– Deep Visual Interpretation and Diagnosis for Image Classifiers (**DeepVID)** is a model-agnostic approach for interpreting and diagnosing images classifiers, providing a rich user interface for understanding convolutional neural networks (CNNs).

– DeepVID is one tool in a suite of tools being developed for understanding AI models.

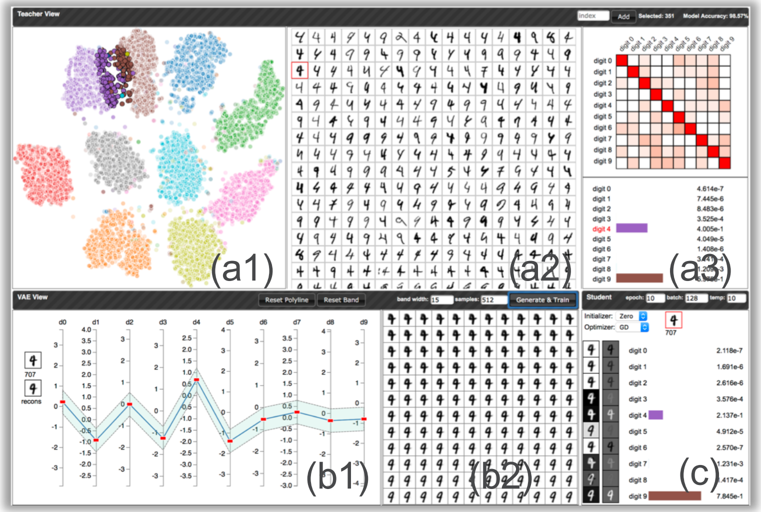


Figure: DeepVID is a visual analytics interface for understanding an image classifier based on variational autodecoder (VAE). Our goal is to understand what knowledge the neural network has acquired enabling it to perform the image classification tasks. We visualize the various aspects of the neural models that will help the developer to optimize and diagnose the classification model.

Extreme-scale Distribution-based Data Analysis

Robust I/O Performance Modeling by Automated Hardware/Software Change Detection

Scientific Achievement

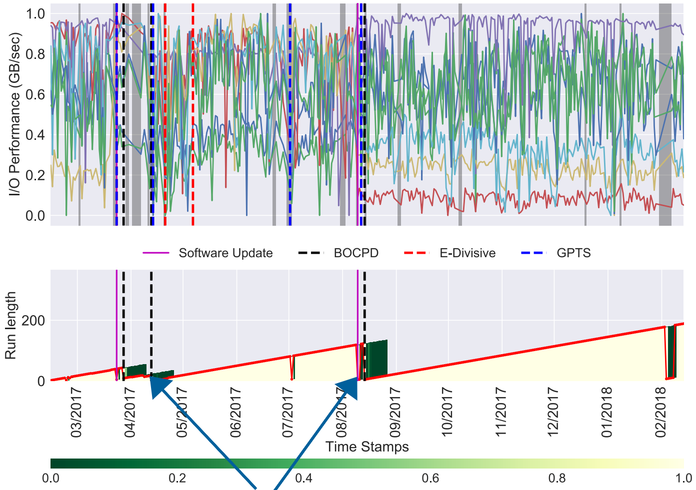
Developed a machine-learning-based I/O performance modeling approach that is robust to HPC system state changes (e.g., hardware degradation, hardware replacement, software upgrades).

Significance and Impact

Automatically identifies hardware and software changes that affect I/O performance in HPC systems and adapts our performance model, allowing better prediction and potentially improving the system utilization and application scheduling.

Research Details

- Online Bayesian detection to automatically identify the location of events that lead to changes in near-real time
- Moment-matching transformation that converts the training data collected before the change to be useful for retraining.
- Approach demonstrated on I/O performance data obtained on Lustre file system at NERSC.



Online method that monitors the change in the I/O performance of an application and adapt the model to these changes

We use application I/O performance data collected on Cori, a production supercomputing system at NERSC, to demonstrate the effectiveness of our approach. The results show that our robust models obtain ***significant reduction in prediction error---from 20.13% to 8.28%*** when the proposed approaches were used in I/O performance modeling.

S. Madireddy,et al. Adaptive Learning for Concept Drift in Application Performance Modeling, Preprint, ANL/MCS-P9132-0918, 2019.

Application Engagement

ANSHU DUBEY
ANL

SAM WILLIAMS
LBNL

Gotta Catch'em All!

Title	PI	Prog.	RAPIDS Member(s)
Coupling Approaches for Next-Generation Architectures (CANGA)	P. Jones	BER	Peterka
Prob. Sea-Level Projections from Ice Sheet and Earth System	S. Price	BER	Patchett
An integrated system for optimization of sensor networks	D. Ricciuto	BER	Steed, Klasky, Podhorszki
Advancing Catalysis Modeling	M. H. Gordon	BES	Williams, Ibrahim
Comp. Framework for Unbiased Studies of Correlated Electron	T. Maier	BES	Huck
AToM: Advanced Tokamak Modeling Environment	J. Candy	FES	Bernholdt
Plasma Surface Interactions (PSI-2)	B. Wirth	FES	Bernholdt, Roth, Pugmire,
Center for Tokamak Transients Simulations (CTTS)	S. Jardin	FES	Williams
Integrated Simulation of Energetic Particles in Plasmas (ISEP)	Z. Lin	FES	Williams, Klasky, Pugmire
Multiscale Gyrokinetic Turbulence (MGK)	D. Hatch	FES	Shan
High-fidelity Boundary Plasma Simulation (HBPS)	C. S. Chang	FES	Klasky, Podhorszki
Tokamak Disruption Simulation	X. Tang	FES	Brugger, Dubey
Inference at Extreme Scale	S. Habib	HEP	Yoo, Morozov, Balaprakash
HEP Data Analytics on HPC	J. Kowalkowski	HEP	Peterka, Ross
HPC Framework for Event Generation at Colliders	S. Hoeche	HEP	Hovland
HEP Event Reconstruction with Cutting Edge Computing	G. Cerati	HEP	Norris, Lee, Vetter
Simulation of Fission Gas in Uranium Oxide Nuclear Fuel	D. Andersson	NE	Bernholdt, Roth
Towards Exascale Astrophysics of Mergers and SuperNova	W. R. Hix	NP	Dubey, Huck
Nuclear Low Energy Initiative (NUCLEI)	J. Carlson	NP	Norris

Thanks to the RAPIDS Team!

… and more …

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program.

RAPIDS

For general questions:

Rob Ross <rross@mcs.anl.gov>

Lenny Oliker <LOliker@lbl.gov>

For engagement discussion:

Anshu Dubey <adubey@anl.gov>

Sam Williams <swwilliams@lbl.gov>

On the web:

http://www.rapids-scidac.org

… or just reach out to the RAPIDS person that you already know!